

AK

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-295247

(43)Date of publication of application : 20.10.2000

(51)Int.Cl.

H04L 12/28

H04L 12/56

(21)Application number : 11-244911

(71)Applicant : NEC CORP

(22)Date of filing : 31.08.1999

(72)Inventor : MARK GOUDORUU
SUTAFUROSU KORIOPAROSU
SATEISHI RAO

(30)Priority

Priority number : 99 128685
99 342976Priority date : 09.04.1999
30.06.1999Priority country : US
US

(54) SCHEDULING METHOD IN QUEUING SYSTEM AND CHUTING BLANKS SWITCH

(57)Abstract:

PROBLEM TO BE SOLVED: To obtain a high stability with a simple operation by discriminating one port having a maximum cell arrival rate and adding a dummy cell to ports except the maximum rate cell arrival port in order to permit the arrival rates with respect to the whole ports to be nearly equal.

SOLUTION: A chuting blanks switch is provided with four input ports 6-11 to 6-14 and four output ports 6-21 to 6-24, for example. The cell arrives at the input ports 6-11 to 6-14 by different rate. A rate discriminator 6.3 discriminates the input/output ports 6-11 to 6-14 and 6-21 to 6-24 of the maximum arrival rate. A filling settler 6.4 settles a dummy flow corresponding to a proper queue to permit the arrival rates with respect to the whole ports 6-11 to 6-14 and 6-21 to 6-24 to be equal. A dummy cell generator 6.5 generates the dummy cell. A scheduler 6.6 executes routing by a heuristic MBM method without considering a queue length.



LEGAL STATUS

[Date of request for examination]

10.07.2000

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's
decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

3175735

[Date of registration]

06.04.2001

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of
rejection]

[Date of extinction of right]

Copyright (C); 1998,2000 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2000-295247

(P2000-295247A)

(43) 公開日 平成12年10月20日 (2000. 10. 20)

(51) Int. Cl.⁹

識別記号

F I

フィコード (参考)

H 0 4 L 12/28

H 0 4 L 11/20

G 5 K 0 3 0

12/56

1 0 2 Z 9 A 0 0 1

審査請求 有 請求項の数18 O L (全 16 頁)

(21) 出願番号 特願平11-244911

(22) 出願日 平成11年8月31日 (1999. 8. 31)

(31) 優先権主張番号 09/342976

(32) 優先日 平成11年6月30日 (1999. 6. 30)

(33) 優先権主張国 米国 (US)

(31) 優先権主張番号 60/128685

(32) 優先日 平成11年4月9日 (1999. 4. 9)

(33) 優先権主張国 米国 (US)

(71) 出願人 000004237

日本電気株式会社

東京都港区芝五丁目7番1号

(72) 発明者 マーク・ゴウドルー

アメリカ合衆国、ニュージャージー

08540 プリンストン、4 インディペン

デンス ウエイ、エヌ・イー・シー・ユ

ー・エス・イー・インク内

(74) 代理人 100097157

弁理士 桂木 雄二

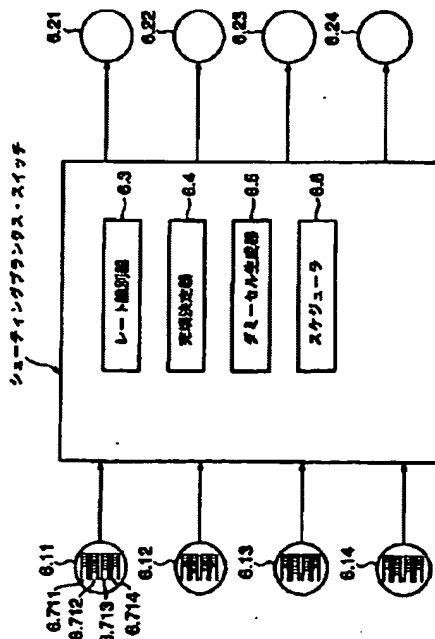
最終頁に続く

(54) 【発明の名称】 キューイングシステムにおけるスケジューリング方法及びシューティングブランクス・スイッチ

(57) 【要約】

【課題】 MWBM (最大重み2部マッチング) スケジューラの優れた安定性を達成するとともに、より単純な MBM (最大サイズ2部マッチング) または MBM ベースのヒューリスティックスを実行するスイッチのためのスケジューリング方法を提供する

【解決手段】 スwitchのM個の入力ポートとN個の出力ポートの間でセルがルーティングされる。セルは、ポートのサブセットごとに異なるレートで到着する。M個の入力ポートはそれぞれN個のキューを有することにより、全部でM×N個のキューが存在する。スケジューリング方法は、まず、1つのポートを、最大のセル到着レートを有するとして識別する。次に、すべてのポートに対する到着レートがほぼ等しくなるように、最大のセル到着レートを有するとして識別されたポート以外のすべてのポートにダミーセルを追加する。



【特許請求の範囲】

【請求項1】 N個の入力キューを1つのグループとしてM個のグループを有し、前記N個の入力キューはそれぞれ、N個のサービスのうちの相異なる1つを要求し、M及びNは自然数であり、従って前記システムはM×N個のキューを有するシステムのためのスケジューリング方法において、

(a) M個のグループ及びN個のサービスに対するエントリの到着レートを判定するステップと、

(b) 前記M個のグループ及びN個のサービスのうち最大到着レートを有する1つを識別するステップと、

(c) 前記M個のグループ及びN個のサービスのうち、前記ステップ(b)で識別されたグループまたはサービス以外のすべてのグループ及びサービスに対して、前記M個のグループ及びN個のサービスに対するエントリの到着レートがほぼ等しくなるように、ダミーエントリを追加するステップと、

(d) 入力キューの長さを考慮しないスケジューリング法を使用するステップと、

からなることを特徴とするスケジューリング方法。

【請求項2】 M個の入力ポート及びN個の出力ポートを有し、前記M個の入力ポートと前記N個の出力ポートの間でセルがルーティングされ、前記セルはポートのサブセットごとに異なるレートで到着し、前記M個の入力ポートはそれぞれN個のキューを有することにより全部でM×N個のキューが存在するスイッチのルーティング方法において、

(a) 最大のセル到着レートを有する1つのポートを識別するステップと、

(b) すべての入力ポート及び出力ポートに対する到着レートがほぼ等しくなるように、最大のセル到着レートを有するとして識別された入力ポートまたは出力ポート以外のすべての入力ポート及び出力ポートにダミーセルを追加するステップと、

(c) キュー長を考慮しないスケジューリング法を使用するステップとからなり、

前記ステップ(a)において、前記M個の入力ポートのうちの1つの入力ポートに対する到着レートは、当該1つの入力ポートに関連するN個のキューの到着レートの和であり、前記N個の出力ポートのうちの1つの出力ポートに対する到着レートは、前記M個の入力ポートの各入力ポートから当該1つの出力ポートへの到着レートの和である、

ことを特徴とするルーティング方法。

【請求項3】 前記ステップ(b)は、

(b.1) すべてのポートに対する到着レートが等しくなるように適当なキューに対するダミーフローを決定して充填を実行するステップと、

(b.2) 前記ステップ(b.1)で決定された到着レートでダミーセルの生成を実行するステップと、

からなることを特徴とする請求項2記載の方法。

【請求項4】 前記ステップ(b.1)の充填は、

(b.1.1) ポートのうちで最大到着レート Λ のものを識別するステップと、

(b.1.2) 入力ポートiと出力ポートjの間のパケット伝送レート τ_{ij} を、初期伝送レート α_{ij} に初期化するステップと、

(b.1.3) 入力ポートポイント及び出力ポートポイントを1に初期化するステップと、

(b.1.4) 入力ポートiの到着レート R_i 及び出力ポートjの到着レート S_j が Λ より小さくなるまで、前記入力ポートポイント及び前記出力ポートポイントをインクリメントするステップと、

(b.1.5) iとjの間のフローに対するダミーセルを、 R_i が S_j 以上である場合はレート $\beta_{ij} = \Lambda - R_i$ で、それ以外の場合はレート $\beta_{ij} = \Lambda - S_j$ で、追加するステップと、

(b.1.6) 目標到着レート τ_{ij} を $\tau_{ij} + \beta_{ij}$ に更新するステップと、

(b.1.7) すべてのポートについて前記ステップ

(b.1.4)～(b.1.6)を繰り返すステップと、

からなることを特徴とする請求項3記載の方法。

【請求項5】 前記ステップ(b.2)の生成は、目標到着レートが満たされるように各入力ポートにおいてダミーセルを生成し且つ実セルが入力ポートに到着しているときには当該実セルを対応するダミーセルにヒギバックすることによって実行され、

各キューは、該キューにおいて伝送のためにキューイングしているダミーセルの数に対応するクレジット値を保持し、

スケジューラは、前記クレジット値が1以上であるときに、特定のキュー内のセルについて知らされ、当該特定のキューが選択されると前記クレジット値は1だけデクリメントされることを特徴とする請求項3記載の方法。

【請求項6】 ダミーセルは実際に生成されるのではなく、各入力キューが、前記クレジット値に等しい数であってiからjへの伝送のためにキューイングしているダミーセルの数に対応する数 d_{ij} を保持し、

前記数 d_{ij} 及び前記クレジット値は、各時間ステップごとに、確率 τ_{ij} でインクリメントされることを特徴とする請求項5記載の方法。

【請求項7】 iとjの間のコネクションが許可されたときに、実セルがiとjの間の伝送のためにキューイングされている場合には、該実セルが伝送され、 d_{ij} が1だけデクリメントされることを特徴とする請求項6記載の方法。

【請求項8】 ダミーセルは生成されるのではなく、各入力キューが数 c_{ij} を保持し、

前記数 c_{ij} は、前記クレジット値に等しく、

前記数 c_{ij} 及び対応するクレジット値は、 i から j への伝送のためにキューイングされるダミーセルの数に対応する浮動小数点数であり、

前記数 c_{ij} 及び前記クレジット値は、各時間ステップごとに、0と1の間の小数をとる τ_{ij} だけインクリメントされることを特徴とする請求項5記載の方法。

【請求項9】 i と j の間のコネクションが許可されたときに、実セルが i と j の間の伝送のためにキューイングされている場合には、該実セルが伝送されることを特徴とする請求項8記載の方法。

【請求項10】 $c_{ij} \geq 1$ の場合、または、実セルが i と j の間のキューに存在する場合、該キューは、キュー i と j の間でセルを伝送することができることをスケジューラに通知することを特徴とする請求項8記載の方法。

【請求項11】 シューティングブランク・スイッチにおいて、

$M+N$ 個のポートと、レート識別器と、充填決定器と、ダミーセル生成器と、スケジューラと、を有し、

前記ポートは、 M 個の入力ポート及び N 個の出力ポートからなり、該 M 個の入力ポートと該 N 個の出力ポートの間でセルがルーティングされ、該セルは、ポートのサブセットごとに異なるレートで到着し、前記 M 個の入力ポートはそれぞれ N 個のキューを有することにより、全部で $M \times N$ 個のキューが存在し、

前記レート識別器は、最大のセル到着レートを有する1つのポートを識別し、前記 M 個の入力ポートのうちの1つの入力ポートに対する到着レートは、該1つの入力ポートの N 個のキューに対する到着レートの和であり、前記 N 個の出力ポートのうちの1つの出力ポートに対する到着レートは、前記 M 個の入力ポートのそれぞれから該1つの出力ポートへの到着レートの和であり、

前記充填決定器は、すべてのポートに対する到着レートが等しくなるように適当なキューに対するダミーフローを決定し、

前記ダミーセル生成器は、前記充填決定器によって決定された到着レートでダミーセルを生成し、

前記スケジューラはキュー長を考慮しない、

ことを特徴とするシューティングブランク・スイッチ。

【請求項12】 前記生成器は、目標到着レートが満たされるように各入力ポートにおいてダミーセルを生成し、実セルが入力ポートに到着しているときには該実セルを対応するダミーセルにビジーバックし、各キューは、該キューにおいて伝送のためにキューイングされているダミーセルの数に対応するクレジット値を保持し、

前記スケジューラは、前記クレジット値が1以上であるときに、特定のキュー内のセルについて知らされ、該特定のキューが選択されると前記クレジット値は1だけデ

クリメントされる、

ことを特徴とする請求項11記載のシューティングブランク・スイッチ。

【請求項13】 ダミーセルは実際に生成されるのではなく、各入力キューが数 d_{ij} を保持し、前記数 d_{ij} は前記クレジット値に等しく、 i から j への伝送のためにキューイングされるダミーセルの数に対応し、

前記数 d_{ij} は、各時間ステップごとに、目標レート τ_{ij} に等しい確率でインクリメントされる、

ことを特徴とする請求項11記載のシューティングブランク・スイッチ。

【請求項14】 i と j の間のコネクションが許可されたときに、実セルが i と j の間の伝送のためにキューイングされている場合には、該実セルが伝送されることを特徴とする請求項13記載のシューティングブランク・スイッチ。

【請求項15】 ダミーセルは生成されるのではなく、各入力キューは数 c_{ij} を保持し、

前記数 c_{ij} は、前記クレジット値に等しく、

前記数 c_{ij} 及び対応するクレジット値は、 i から j への伝送のためにキューイングされるダミーセルの数に対応する浮動小数点数であり、

前記数 c_{ij} は、各時間ステップごとに、0と1の間の小数をとる τ_{ij} だけインクリメントされる、

ことを特徴とする請求項12記載のシューティングブランク・スイッチ。

【請求項16】 i と j の間のコネクションが許可されたときに、実セルが i と j の間の伝送のためにキューイングされている場合には、該実セルが伝送されることを特徴とする請求項15記載のシューティングブランク・スイッチ。

【請求項17】 $c_{ij} \geq 1$ の場合、または、実セルが i と j の間のキューに存在する場合、前記キューは i と j の間でセルを伝送することができることを前記スケジューラに通知し、 c_{ij} は1だけデクリメントされることを特徴とする請求項16記載のシューティングブランク・スイッチ。

【請求項18】 スwitchのサブセットが請求項11記載のシューティングブランク・スイッチであるようなスイッチを有することを特徴とするネットワークシステム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、複数のサービスを要求する複数のキューを有するキューイングシステムのスケジューリング方法に係り、特に入力キュースイッチにおける高いスループットを達成するためのシューティングブランク(shooting blanks)方法に関する。

【0002】

【従来の技術】 いくつかのアプリケーションは、複数の

サービスを要求する複数のキューの割当てを必要とする。本発明は、そのようなキューイングシステムにおけるスケジューリング技術の改良に関する。

【0003】[グラフ理論の用語] 本発明の理解を容易にするために、ここでいくつかのグラフ理論用語を定義する。

【0004】2部グラフとは、図1に例示するように、複数の頂点を1. 1~1. 4と1. 5~1. 8という2つの集合に分割することができ、辺1. 9がこれらの2つの集合の間にのみ位置するような無向グラフをいう。サービスの集合を要求するキューの集合は、このような2部グラフによって表現することができる。辺は、第1の頂点と第2の頂点の間のキューを表す。2頂点間の辺の存在は、対応するキューが第2の頂点によって表されるサービスを要求することを示す。

【0005】2部マッチングとは、どの頂点も複数の辺の端点にならないような辺の部分集合である。マッチングを求めることは、キューと、要求されるサービスとの間の置換（あるいは並び替え）の1つを選択することに対応する。

【0006】・最大(サイズ)マッチングとは、可能な最大数の辺からなるマッチングである。n及びmを、それぞれ、現在の状態を表す基礎となるグラフの頂点及び辺の数とすると、N個のサービスを要求するN²個のキューを表すN×Nシステムの場合、 $n = 2N$ であり、 $m \leq N^2$ である。

【0007】・最大重みマッチングとは、グラフ内の辺に関連づけられた重みも考慮に入れたマッチングをいう。本発明では、重みは、対応するキューのサイズに等しくする。

【0008】[システムモデル] 背景的知识及び本発明をよりよく理解するため、次に、システムモデルについて説明する。当業者には明らかなように、システムモデルは、ここで説明する事項の理解を簡単にするためにのみ用いられるものであって、本発明の技術的範囲を制限するように解釈してはならない。

【0009】ここでは、N個の入力ポート及びN個の出力ポートを有するクロスバスイッチを含む交換ファブリックを使用するが、本発明は、M個の入力ポート及びN個の出力ポートを有するスイッチにも適用可能である。このクロスバスイッチは、1つの時間ステップ（以下、1ステップという。）で、固定サイズのパケットの任意の置換（並び替え）をルーティングすることができる。1ステップで、入力ポートを出ること、または、出力に到達することが可能なセルは1個だけである。従って、多数個のセルは、入力で待合せ（キューイング）されることがある。入力ポートごとに1個のキューを使用すると、HOL(head-of-line)ブロッキング効果を引き起こし、スループットをひどく制限する可能性がある。このようなブロッキング効果について詳細な情報は、Mark

J. Karol, Michael G. Hluchyj, and Samuel P. Morgan, "Input versus output queuing on a space-division packet switch", IEEE Transactions on Communications, COM-35(12):1347-1356, December 1987, に記載されている。

【0010】さまざまな研究グループが、入力ポートごとにN個のキュー（N個の出力ポートのそれぞれに対して1個ずつのキュー）を使用することを提唱している。例えば、

・Thomas E. Anderson, Susan S. Owicki, James B. Saxe, and Charles P. Thacker, "High-speed switch scheduling for local-area networks", ACM Transactions on Computer Systems, 11(4):319-352, November 1993

・M. Karol, K. Eng, and H. Obara, "Improving the performance of input-queued ATM packet-switching", In Proceedings IEEE INFOCOM'92, p.110-115, 1992

・Nicholas William McKeown, "Scheduling Algorithms for Input-Queued Cell Switches", PhD thesis, University of California at Berkeley, 1995

を参照のこと。

【0011】これらに記載されている技術（仮想出力キューイング(VOQ: virtual output queuing)という。）は、背景的知识として、本発明の好ましい実施例を説明するために用いられるモデルである。直観的には、VOQによれば、スイッチは、置換を選択する際に優れたフレキシビリティを有する。図2は、N=4のVOQスイッチの一例を示す模式的構成図である。

【0012】入力ポートiから出力ポートjに行く必要があるパケットの到着レートを α_{ij} で表す。レート50%とは、平均として、1ステップおきに1個のパケットが、入力ポートiから出力ポートjに送られることを意味する。パケットが入力ポートiに到着するレートを R_i とする。これは $\sum_j \alpha_{ij}$ に等しい。パケットが出力ポートjに到着するレートを S_j とする。これは $\sum_i \alpha_{ij}$ に等しい。なお、この総和において、iはすべての入力ポートを表し、jはすべての出力ポートを表す。正式な用語ではないが、無限なく成長する入力ポートがない場合にスイッチは安定であるという。すべてのi、jについて、 $R_i, S_j < 100\%$ のとき、トラフィックは許容されるという。

【0013】[ルータスイッチとスケジューリング] このようなキューイングシステムの重要な例は、インターネットのようなアプリケーションで用いられるルータスイッチである。本発明は、2部グラフを用いて表現可能なキューイングシステムの重要な例としてスイッチを使用する。ただし、本発明は、スイッチに限定されるものではなく、2部グラフを用いて表現可能な任意のキューイングシステムに適用可能であることに注意すべきである。

【0014】インターネットの発展は、ルータ及びスイ

ッチの需要をますます増大させ続けている。毎秒数十ないし数百ギガビットの広帯域を処理可能で、それとともに多数の入出力ポートを有するスイッチ設計法が現在では利用可能である。ネットワークシステムにおいて、セル（あるいはパケット）は、いくつかの入力ポートを通じてスイッチに到着する。これらのセルは、スイッチによって使用されるスケジューリング方法に基づいて、1つまたはいくつかの出力ポートを通じてルーティングされる。なお、本明細書では、パケットとセルという用語は交換可能であり、本発明は、いかなる種類のセルあるいはパケットにも制限されるものではない。

【0015】広帯域スイッチの設計者は、交換ファブリックを通じてセルをスケジューリングする有効な技術を提供するという問題に直面している。現在、理論的な観点から、知られている最良のアプローチは、従来の最大重み2部マッチング(MWBM: Maximum Weight Bipartite Matching)を利用するものである。入出力ポートが100%容量で動作する必要がない限り、かなり一般的な確率的仮定のもとで、MWBMは安定であることが、Tassiulas and Ephremidesと、McKeown, Anantharam, and Walrandとによって独立に示された。(入力ポートまたは出力ポートに関連する)いずれのキューも無制限に成長しないとき、スケジューリング方法は安定であるという。MWBMに関するその他の情報については、

・Leandros Tassiulas and Anthony Ephremides, "Stability properties of constrained queuing systems and scheduling policies for maximum throughput in multihop radio networks", IEEE Transactions on Automatic Control, 37(12):1936-1948, December 1992

・Nick McKeown, Venkat Anantharam, and Jean Walrand, "Achieving 100% throughput in an input-queued switch", In Proceedings IEEE INFOCOM' 96, p.296-302, San Francisco, CA, March 1996

に記載されている。入出力ポートの負荷がフルでない場合のトラフィックパターンは許容されるとみなされる。

【0016】しかしながら、MWBMの既知の最良の方法の実行時間は、 $O(N^2 \log N + Nm)$ 及び $O(N^{1/2} m \log NC)$ である。ただし、 N はスイッチ内の入力ポートの数であり、 $m \leq N^2$ は空でないキューの数であり、 C は最大キューサイズである。MWBM法の実行時間に関する詳細な情報については、

・M. L. Friedman and R. E. Tarjan, "Fibonacci heaps and their uses in improved network optimization algorithms", Journal of the ACM, 34:596-615, 1987

・H. N. Gabow and R. E. Tarjan, "Faster scaling algorithms for network problems", SIAM Journal on Computing, 18:1013-1036, 1989

に記載されている。さらに、これらの方法は本質的にシークンシャルである。置換ごとに伴う実行時間のために、MWBMの完全な実現は非実証的である。MWBM

の完全な実現は、セル時間あたり1つの置換計算を必要とする広帯域スイッチでは可能性はあまりないと考えられている。置換はマッチングと等価である。

【0017】これに対して、最大(サイズ)2部マッチング(MBM: Maximum (Size) Bipartite Matching)に基づくスケジューリングアプローチには次の2つの主要な利点がある。

【0018】(1) 第1に、最良の既知のMBM法の実行時間は $O(N^{1/2} m)$ である。MBMの実行時間に関するさらに詳細な情報については、

・J. E. Hopcroft and R. M. Karp, "An $n^{5/2}$ algorithm for maximum matching in bipartite graphs", SIAM Journal on Computing, 2:225-231, 1973

・E. A. Dinitz, "Algorithm for solution of a problem of maximum flow in networks with power estimation", Soviet Math. Dokl., 11:1277-1280, 1970

・S. Even and R. E. Tarjan, "Network flow and testing graph connectivity", SIAM Journal on Computing, 4:507-518, 1975

に記載されている。

【0019】(2) 第2に、この方法の入力は、各キューが空であるか否かについての知識(N^2 ビット)しか必要としない。これに対して、MWBM法は、各キューの長さについての知識(N^2 個の整数。ただし、整数のサイズには上限がない。)を必要とする。

【0020】しかしながら、MWBMと比べた場合のMBMの欠点は、MBMは、すべての許容トラフィックパターンに対して安定であるとは限らないことである。特に、McKeown et al.は、許容されるがMBMのもとで安定ではない $N=3$ の単純なトラフィックパターンを示している(Nick McKeown, Venkat Anantharam, and Jean Walrand, "Achieving 100% throughput in an input-queued switch", In Proceedings IEEE INFOCOM' 96, p.296-302, San Francisco, CA, を参照)。

【0021】しかし、MBMが安定になるような許容トラフィックパターンの興味深いクラスがあることが予想される。このクラスのトラフィックは、平衡(balanced)トラフィックパターンである。平衡トラフィックパターンの場合、すべての入出力ポートの負荷は均一である。以下の事実は、上記の予想を正当化する。

【0022】1. MBMは、平衡スケジューリング問題のオフラインバージョンに対して100%の利用率を提供することが知られている。各入力ポートがちょうど T 個のセルを送出することを必要とし、各出力ポートがちょうど T 個のセルを受け取ることを必要としている場合、 NT 個のすべてのセルをちょうど T 回の置換でルーティングすることが可能である。以下を参照のこと：

・G. Birkhoff, "Tres observaciones sobre el algebra lineal", Univ. Nac. Tucuman Rev. Ser. A5, p.147-150, 1946;

Shizhao Li and Nirwan Ansari, "Input-queued switching with QoS guarantees", In Proceedings IEEE INFOCOM'99, volume 3, p.1152-1159, March 1999). 従って、MBM法をT回使用して置換を選択することができる。

【0023】2. 現在、MBMアプローチの既知のすべてのシミュレーションは、トラフィックフローが平衡であるときは必ず100%負荷まで安定である。シミュレーションは、不平衡フローの場合にのみ、不安定性を示している。

【0024】置換選択問題は、グラフ理論用語でモデル化することができる。2部グラフとは、図1に示すように、頂点を2つの集合に分割することが可能であり、各辺がこれらの2つの集合間にあるような無向グラフである。2部グラフは、スイッチを表現するために用いることができる。この場合、各入力ポートは一方の集合のノード(節点、すなわち頂点)によって表され、各出力ポートは、他方の集合のノードによって表される。2頂点間に辺が存在することは、スイッチ内の対応する入力ポートと出力ポートとの間でルーティングされるべきセルが存在することを示す。極大マッチングとは、マッチングを維持したままでは追加することができないようなマッチングのことである。

【0025】文献に現れているほとんどのスケジューリング方法は、2つのカテゴリに分けられる。キューのサイズを考慮するもの(MWBMによって例示される)と、考慮しないもの(MBMによって例示される)である。

【0026】MWBMアプローチでは、2部グラフの辺に重み値が割り当てられる。一般に、入力ポート*i*から出力ポート*j*への辺には、入力ポート*i*にキューイングされた、出力ポート*j*を宛先とするパケットの数 q_{ij} に等しい重みが割り当てられる。その後、得られたグラフにおいて最大重みマッチングを計算する。前述のように、MWBMアプローチは、トラフィックパターンが不平衡のときでも安定性を備える。しかし、既知の最良の方法の実行時間は、 $O(N^2 \log N + Nm)$ 及び $O(N^{1/2} m \log NC)$ である。さらに、これらの方法は本質的にシーケンシャル(直列的、逐次的)であり、良好なハードウェア実現が困難である。MWBMを近似しようとするヒューリスティクスも提案されているが、それらもまた、シーケンシャル計算のボトルネックによって制限されるようである。以下参照のこと:

Adisak Mekittikul and Nick McKeown, "A practical scheduling method to achieve 100% throughput in input-queued switches", In Proceedings IEEE INFOCOM'98, p.792-799, San Francisco, CA, March 1998;

Anthony C. Kam and Kai-Young Siu, "Linear complexity methods for bandwidth reservations and delay guarantees in input-queued switches with no speedu

p", IEEE International Conference on Network Protocols, Austin, Texas, October 1998).

【0027】MBMアプローチでは、可能な最大数のパケットがタイムスロットごとにルーティングされる。MWBM法よりも漸近的に優れているが(既知の最良のMBM法の実行時間は $O(N^{1/2} m)$ でしかない)、MBMは依然として、広帯域スイッチの速度制約のもとでは実現が困難な方法である。PIMやSLIPのようないくつかのヒューリスティック近似が開発されている。これらは、より簡単な実現が可能であり、最大サイズマッチングに近似しようとするものである。以下参照のこと:

Nick McKeown, Martin Izzard, Adisak Mekittikul, William Ellersick, and Mark Horowitz, "The Tiny Tern: A packet core switch", In Hot Interconnects IV, p.161-173, Stanford University, August 1996;

Nicholas William McKeown, "Scheduling Algorithms for Input-Queued Cell Switches", PhD thesis, University of California at Berkeley, 1995;

Thomas E. Anderson, Susan S. Owicki, James B. Saxe, and Charles P. Thacker, "High-speed switch scheduling for local-area networks", ACM Transactions on Computer Systems, 11(4):319-352, November 1993)。これらのヒューリスティックアプローチもまた、不平衡の許容トラフィックパターンに対しては不安定性を示すことがある。

【0028】[MBMに対するスケジューリングヒューリスティクス] 当業者には理解されるように、MBM法についていくつかのヒューリスティクスが使用可能である。ここで説明するMBM関連のヒューリスティクスは、後述するセクションIV Dに記載する実験で、さまざまなMBM法の包括的な比較研究のために用いられる。

【0029】PIM: 並列反復マッチング(Parallel Iterative Matching)は、平均として $O(\log n)$ で極大マッチングに収束するヒューリスティクスである(Thomas E. Anderson, Susan S. Owicki, James B. Saxe, and Charles P. Thacker, "High-speed switch scheduling for local-area networks", ACM Transactions on Computer Systems, 11(4):319-352, November 1993, を参照)。各イテレーションごとに、マッチングしていない入力、キューイングされているセルを有するあらゆる出力に対して提案(要求)をする。マッチングしていない出力は、要求を受け取った場合、すべての要求にわたって一様ランダムに選択を行うことによって1つの要求を受け入れる。1つの入力複数の許可を受け取った場合、その入力は、それらのうちからランダムに1つを選択する。

【0030】RND: RNDヒューリスティクスはPIMに類似している。相違点は、マッチングしていな

いあらゆる入力 i は、 i がキューイングされているセルを有するようなランダムに選択された出力ポート j に対する 1 つの要求のみを行う点である。

【0031】RR-RND: ラウンドロビン法 RR-RND では、与えられたタイムスロット t に対するスケジューリングは、固定された優先スケジュールによって決定される。最高の優先度の入力 i_1 は、それがキューイングされたセルを有するような出力のうちからランダムに選択を行う。次に高い優先度の入力 i_2 は、取られていない出力のうちから次の選択を行い、以下同様である。次のタイムスロット $t+1$ において、入力ポート優先度は、 N を法として 1 だけインクリメントされる。この優先方式の利点は、パイプライン機構を有するハードウェアで容易に実現可能なことである。

【0032】SLIP: SLIP は、McKeown によって開発され、広範に解析されている (Nicholas William McKeown, "Scheduling Algorithms for Input-Queued Cell Switches", PhD thesis, University of California at Berkeley, 1995, を参照)。各入力ポート及び出力ポートは、別々の優先度ホイール (回転盤) を有する。1 つの入力は、それがキューイングされたセルを有するようなあらゆる出力に 1 つずつ要求を送る。1 つの出力 j は、自己の優先度ホイールに基づいて、ポート i からの要求を受け入れることを選択する。 i によって受け入れられた場合、 j における優先度ホイールは、 $(i+1) \bmod N$ にセットされる。入力ポートは、優先度ホイールに基づいて許可を受け入れ、自己のホイールを同様に回す。

【0033】SHAKEUP: SHAKEUP 法は、米国特許出願第 09/342,975 号 (本特許出願と同日出願の特願平 11-244648 号) において本発明の発明者によって導入されたものである。SHAKEUP は、それ自体でも、他の方法と組み合わせても使用可能であり、大きなパフォーマンス利得を達成することが示されている (前記特願平 11-244648 号を参照のこと)。この方法は、2 部グラフで、何らかの初期マッチングを仮定する。初期マッチングは、例えば、他のヒューリスティクスにより生成される。基本的な考え方は、A 内のマッチングしていない各頂点は、たとえ既存のマッチングを除去することになるとしても、その頂点自身がマッチングを強制されるということである。具体的には、少なくとも 1 本の辺の端点である。集合 A 内のマッチングしていない各頂点 x は、端点となっている辺のうちの 1 つ (例えば、 (x, y)) を一様ランダムに選択する。 $y \in B$ がマッチングしていない場合、 (x, y) をマッチングに追加する。 y が既に辺 (z, y) 、 $z \in A$ によりマッチングしている場合、 (z, y) をマッチングから除去し、 (x, y) を追加する。最後に、SHAKEUP 中に、A 内の複数の頂点が B 内の同じ頂点とマッチングしようとする場合、A 内の競合

する頂点のうちの 1 つをランダムに勝者として選択する。ここで、SHAKEUP は、並列に実現されておらず、A 内のマッチングしていない頂点についてランダムに選択された順序で実行されると仮定する。SHAKEUP は、少なくとも初期マッチングと同じ大きさのマッチングを返す (SHAKEUP に関する詳細な実験及び理論的正当化については、前記米国特許出願及び特願平 11-244648 号を参照)。

【0034】なお、PIM、RND 及び SLIP のヒューリスティクスは、反復ヒューリスティクスのクラスに入る。反復ヒューリスティクスは、1 回の実行 (イテレーション) で極大マッチングを見つけることは保証されない。従って、複数回反復することにより、置換のサイズが増大する可能性がある。

【0035】

【発明が解決しようとする課題】前述のように、MBM 及び MWBM のアプローチはいずれもそれぞれの欠点を有する。従って、MWBM の安定性を有するとともに、MBM のような低い実行時間要求条件を有するスケジューリング方法が、特に広帯域スイッチを実現するために所望される。

【0036】本発明の目的は、一般的なトラフィックパターンに対して高いパフォーマンスのスケジューリング方法を提供することにある。

【0037】特に、本発明の目的は、MWBM スケジューラの優れた安定性を達成すると共に、より単純な MBM または MBM ベースのヒューリスティクスを実行する、スイッチのスケジューリング方法を提供することにある。

【0038】

【課題を解決するための手段】本発明のアプローチは、シューティングブランクスと呼ばれる。その主な考え方について以下で説明する。シューティングブランクス法では、すべての入出力が同じ到着レート λ になるように、ダミーパケットを生成することによって軽い負荷のポートを通るトラフィックを増大させる。ダミーパケット生成をした後、MBM または MBM 関連のヒューリスティクスを適用して、ルーティングされるべき置換を計算する。軽い負荷のポートはダミートラフィックに対応する辺の端点に相当するため、この方法は、いわば隔された形で、これらをマッチングに含めるように選択する。

【0039】本発明は、コンピュータ実装されたスケジューリング方法として、ルーティングに用いられる方法として、シューティングブランクス・スイッチとして、またシューティングブランクス・スイッチを用いたネットワークシステムとして、実現される。

【0040】本発明によれば、 N 個の入力キューを 1 つのグループとして M 個のグループを有するシステムのための、コンピュータにより実現されたスケジューリング

方法が提供される。前記N個の入力キューはそれぞれ、N個のサービスのうちの相異なる1つを要求する。M及びNは自然数である。従って、システムは、 $M \times N$ 個のキューを有することになる。前記スケジューリング方法は、M個のグループ及びN個のサービスに対するエントリの到着レートを判定するステップと、前記M個のグループ及びN個のサービスのうち最大到着レートを有する1つを識別するステップと、前記M個のグループ及びN個のサービスのうち、識別されたグループまたはサービス以外のすべてのグループ及びサービスに、前記M個のグループ及びN個のサービスに対するエントリの到着レートがほぼ等しくなるように、ダミーエントリを追加するステップと、入力キューの長さを考慮しないスケジューリング法を使用するステップとからなる。

【0041】本発明の他の側面によれば、ポートを有するスイッチにおけるルーティング方法が提供される。前記ポートは、M個の入力ポート及びN個の出力ポートのうちの1つであり、該M個の入力ポートと該N個の出力ポートの間でセルがルーティングされる。前記セルは、ポートのサブセットごとに異なるレートで到着する。前記M個の入力ポートはそれぞれN個のキューを有することにより、全部で $M \times N$ 個のキューが存在する。前記ルーティング方法は、1つのポートを、最大のセル到着レートを有するとして識別するステップと、すべての入力ポート及び出力ポートに対する到着レートがほぼ等しくなるように、最大のセル到着レートを有するとして識別された入力ポートまたは出力ポート以外のすべての入力ポート及び出力ポートにダミーセルを追加するステップと、キュー長を考慮しないスケジューリング法を使用するステップとからなり、前記M個の入力ポートのうちの1つの入力ポートに対する到着レートは、該1つの入力ポートのN個のキューに対する到着レートの和であり、前記N個の出力ポートのうちの1つの出力ポートに対する到着レートは、前記M個の入力ポートのそれぞれから該1つの出力ポートへの到着レートの和である。

【0042】好ましくは、ダミーセルは、すべてのポートに対する到着レートが等しくなるように適当なキューに対するダミーフローを決定して充填するステップと、決定された到着レートでダミーセルを生成するステップと、からなるプロセスを使用して追加される。

【0043】好ましくは、充填は、ポートのうちの最大到着レート Λ のものを識別するステップと、入力ポートiと出力ポートjの間のバケット伝送レート α_{ij} を、初期伝送レート α_{ij} に初期化するステップと、入力ポートポイント及び出力ポートポイントを1に初期化するステップと、入力ポートiの到着レート R_i 及び出力ポートjの到着レート S_j が Λ より小さくなるまで、前記入力ポートポイント及び前記出力ポートポイントをインクリメントするステップと、iとjの間のフローに対するダミーセルを、 R_i が S_j 以上である場合はレート $\beta_{ij} = \Lambda$

$-R_i$ で、それ以外の場合は $\beta_{ij} = \Lambda - S_j$ で、追加するステップと、目標到着レート τ_{ij} を $\alpha_{ij} + \beta_{ij}$ に更新するステップと、すべてのポートについて上記ステップを繰り返すステップと、からなるプロセスを使用して実行される。

【0044】好ましくは、生成は、目標到着レートが満たされるように各入力ポートにおいてダミーセルを生成するステップと、実セルが入力ポートに到着しているときには該実セルを対応するダミーセルにヒギバックするステップと、によって実行され、各キューは、該キューにおいて伝送のためにキューイングされているダミーセルの数に対応するクレジット値を保持し、スケジューラは、前記クレジット値が1以上であるときに、特定のキュー内のセルについて知らされ、該特定のキューが選択されると前記クレジット値は1だけデクリメントされる。

【0045】さらに好ましくは、ダミーセルは実際に生成されるのではなく、各入力キューが数 d_{ij} を保持する。この数は、前記クレジット値に等しく、iからjへの伝送のためにキューイングされるダミーセルの数に対応する。前記数 d_{ij} 及び前記クレジット値は、各時間ステップごとに、確率 τ_{ij} でインクリメントされる。

【0046】さらに好ましくは、iとjの間のコネクションが許可されたときに、実セルがiとjの間の伝送のためにキューイングされている場合には、該実セルが伝送される。

【0047】もう1つの改良において、好ましくは、ダミーセルは生成されるのではなく、各入力キューが数 c_{ij} を保持する。この数は、前記クレジット値に等しい。 c_{ij} 及び前記クレジット値は、iからjへの伝送のためにキューイングされるダミーセルの数に対応する浮動小数点数である。前記数 c_{ij} 及び前記クレジット値は、各時間ステップごとに、0と1の間の小数をとる τ_{ij} だけインクリメントされる。

【0048】好ましくは、iとjの間のコネクションが許可されたときに、実セルがiとjの間の伝送のためにキューイングされている場合には、該実セルが伝送される。

【0049】別の改良において、好ましくは、 $c_{ij} \geq 1$ の場合、または、実セルがiとjの間のキューに存在する場合、該キューは、iとjの間でセルを伝送することができるとスケジューラに通知する。

【0050】本発明のもう1つの側面であるシューティングブランクス・スイッチは、 $M+N$ 個のポートと、レート識別器と、充填決定器と、ダミーセル生成器と、スケジューラとを有する。前記ポートは、M個の入力ポート及びN個の出力ポートからなり、該M個の入力ポートと該N個の出力ポートの間でセルがルーティングされる。前記セルは、ポートのサブセットごとに異なるレートで到着する。前記M個の入力ポートはそれぞれN個の

キューを有することにより、全部で $M \times N$ 個のキューが存在する。前記レート識別器は、1つのポートを、最大のセル到着レートに有するとして識別する。前記 M 個の入力ポートのうちの1つの入力ポートに対する到着レートは、該1つの入力ポートの N 個のキューに対する到着レートの和であり、前記 N 個の出力ポートのうちの1つの出力ポートに対する到着レートは、前記 M 個の入力ポートのそれぞれから該1つの出力ポートへの到着レートの和である。前記充填決定器は、すべてのポートに対する到着レートが等しくなるように適当なキューに対するダミーフローを決定する。前記ダミーセル生成器は、前記充填決定器によって決定された到着レートでダミーセルを生成する。前記スケジューラは、キュー長を考慮しない。

【0051】好ましくは、前記生成器は、目標到着レートが満たされるように各入力ポートにおいてダミーセルを生成し、実セルが入力ポートに到着しているときには該実セルを対応するダミーセルにヒギバックする。各キューは、該キューにおいて伝送のためにキューイングされているダミーセルの数に対応するクレジット値を保持し、スケジューラは、前記クレジット値が1以上であるときに、特定のキュー内のセルについて知らされ、該特定のキューが選択されると前記クレジット値は1だけデクリメントされる。

【0052】好ましくは、ダミーセルは実際に生成されるのではなく、各入力キューが数 d_{ij} を保持する。この数は、前記クレジット値に等しく、 i から j への伝送のためにキューイングされるダミーセルの数に対応する。前記数 d_{ij} は、各時間ステップごとに、目標レート r_{ij} に等しい確率でインクリメントされる。

【0053】さらに好ましくは、 i と j の間のコネクションが許可されたときに、実セルが i と j の間の伝送のためにキューイングされている場合には、該実セルが伝送される。

【0054】もう1つの改良において、好ましくは、ダミーセルは生成されるのではなく、各入力キューは数 c_{ij} を保持する。この数は、前記クレジット値に等しい。 c_{ij} 及び対応するクレジット値は、 i から j への伝送のためにキューイングされるダミーセルの数に対応する浮動小数点数である。前記数 c_{ij} は、各時間ステップごとに、0と1の間的小数をとる r_{ij} だけインクリメントされる。

【0055】さらに好ましくは、 i と j の間のコネクションが許可されたときに、実セルが i と j の間の伝送のためにキューイングされている場合には、該実セルが伝送される。

【0056】別の改良において、好ましくは、 $c_{ij} \geq 1$ の場合、または、実セルが i と j の間のキューに存在する場合、該キューは、キュー i と j の間でセルを伝送することができることをスケジューラに通知する。

【0057】本発明のさらにもう1つの側面によれば、スイッチのサブセットがシューティングブランクス・スイッチであるようなスイッチを有するネットワークシステムが提供される。

【0058】

【発明の実施の形態】[IVA. シューティングブランクス・アプローチ] 好ましい実施例は、最も重い負荷のポート（入力または出力のいずれでもよい）において到着レートを Λ に等しくするフローの集合を提供するためのスイッチを使用する。このような場合、パケットは Λ に基づく平均値で到着する。

【0059】図3は、不平衡フローのスイッチの一例を示す模式図である。ここでは、最も重い負荷のポートはいちばん上の入力ポートである。シューティングブランクス法では、すべての入出力が同じ到着レート Λ になるようにダミーパケットを生成することによって、軽い負荷のポートを通るトラフィックを増大させる。ダミーパケット生成をした後、MBMまたはMBM関連のヒュリスティックスを適用してルーティングされるべき置換を計算する。軽負荷のポートはダミートラフィックに対応する辺の端点に相当することが多いため、この方法は、マッチングに含めるようにこれらのダミーパケットを選択するようだまされることになる。

【0060】シューティングブランクス法は、3つの主要な利点を有する。(i) MBM関連の方法を使用していながら、MWBMの優れた安定性に匹敵することが実験的に示される（セクションIVDを参照）。(ii) MBMヒュリスティックスの上にシューティングブランクスの実現することによるオーバーヘッドは小さく（セクションIVBを参照）、MWBMあるいはMBMの複雑さに比べても確実に無視できる。(iii) このアプローチはモジュール性がある。すなわち、このアプローチは、トラフィックを「前処理」することに集中しており、任意のスケジューリング方法と組み合わせることができる。

【0061】[IVB. シューティングブランクスの実現] 好ましい実施例の実現の詳細について以下説明する。好ましい実施例のスイッチは、サービス品質(QoS)環境で動作し、フローは、許可された帯域を決定する契約を有する。この帯域はレートで表され、 $i-j$ リンク、すなわち、入力ポート i と出力ポート j の間のリンクのレート α_{ij} は、そのリンク上のすべての顧客のために予約されたレートの和である。契約で合意されたレートは呼受付コントローラ(CAC: Call Admission Controller)には既知である。ネットワークシステムは、どのフローも、それに割り当てられたレートを超過しないことを保証する。レートは、大きい時間ウィンドウを通じて一定であることを仮定することができるくらいまれにしか変化しない。

【0062】本発明のシューティングブランクス法は、

3個のコンポーネントを有する。

【0063】(i) 充填 (Fill-Up) コンポーネントでは、すべての入力ポート及び出力ポートに対するレート R_i 、 S_j が等しいことを保証するように、適当な入出力ペアに対してダミーフローを決定する。

【0064】(ii) 生成 (Generation) コンポーネントは、充填コンポーネントによって決定されたレートでダミーパケットを生成することを担当する。

【0065】(iii) 実際のスケジューリングコンポーネントでは、選択された任意の方法（好ましくは、MBMまたはMBM関連のヒュースティックス）を実行して、置換を計算することができる。

【0066】[IVB. 1 充填コンポーネント] 図4は、図3のトラフィックに対する充填の効果の一例を示す模式図である。充填コンポーネントは、グリーディ（欲張り）手順を用いて実現される。以下、 Λ を入出力のレートのうちの最大値とする。このコンポーネントは新しいレート r_{ij} を出力しようとする。 r_{ij} はオリジナルレート値 α_{ij} に初期化される。グリーディ手順は入出力ポートを1から N への順に進む。2つのポイントが設けられ、一方は現在の入力 i であり、他方は現在の出力 j である。これらのポイントは、 Λ より小さい負荷の入力ポート及び出力ポートが見つかるまでインクリメントされる。 i 及び j の両方に対して、 R_i 、 $S_j < \Lambda$ である。 $\Lambda - R_i < \Lambda - S_j$ と仮定しても一般性を失わない。レート $\beta_{ij} = \Lambda - R_i$ の i から j へのフローを追加し、レート r_{ij} を $r_{ij} + \beta_{ij}$ に更新する。入力ポイントを、 $R_i < \Lambda$ であるような次の $i' > i$ に移動する。このステップを繰り返す。これで、充填コンポーネントの記述は完了する。

【0067】充填コンポーネントは、高々 $2N-1$ 個のフローを追加することが分かる。各ステップで、一樣な負荷 Λ を達成するという目標に向かって進むように辺を追加することが常に可能である。これは、入力レートの和が常に出力レートの和に等しいということによる。従って、この方法は正しい。

【0068】[IVB. 2 生成コンポーネント] 生成コンポーネントの実現について説明する。最も重い負荷のポートのレートを Λ とする。このコンポーネントは、充填コンポーネントによって決定されたレートでダミーフローを生成しなければならない。さらに複雑なことは、実フロー（例えば、入力 i から）がその契約で合意したレート以下である場合、本発明の方法は、充填コンポーネントによって i に対して設定されたレート Λ を満たすように、ダミーパケットが生成されるレートを増大させなければならない。

【0069】生成コンポーネントに対する上記の2つの要件に効率的に対処するために、「ヒギバック」方式を用いる。入力 i は、ちょうど Λ の全レートで、充填コンポーネントによって決定された実フロー及びダミーフ

ローの宛先へ向かうダミーパケットを生成する。すべてのスケジューラは、コネクション $i-j$ に対して、ダミーパケットの有無に対応して1または0とみる。実パケットが入力 i において宛先を j としてキューイングされた場合、これは、次にスケジューラが $i-j$ コネクションを許可するときに、ダミーパケットの先頭に移される。ヒギバック方式によれば、シューティングブランクスの実現は、実フローの高価なポリシング (policing) をさらに行うことなしに、しかも、任意のリンクにおいてダミーパケット生成に固定レートを使用しながら（これはハードウェアで容易に実現可能な方法である。）達成される。

【0070】さらに効率を改善するため、さらに好ましい実施例では、ダミーパケットは、明示的に生成され保持されるのではない。入出力ポートのあらゆるペア (i, j) に対して、 i から j への伝送のためにキューイングされているダミーパケットの数 d_{ij} のみが保持される。この数は、各時間ステップごとに、確率 r_{ij} でインクリメントされる。 $d_{ij} > 0$ の場合、スケジューラは1（すなわち、 $i-j$ 辺が2部グラフ内に存在する）と判断し、そうでない場合は0と判断する。スケジューラが $i-j$ コネクション（これは $d_{ij} > 0$ を意味する。）を許可し、 $q_{ij} > 0$ （これは、伝送のためにキューイングされている実パケットがあることを意味する。）である場合、実パケットがルーティングされ、 d_{ij} は1だけデクリメントされる。 $i-j$ コネクションが許可され、実パケットがキューイングされていない場合、単に、 d_{ij} が1だけデクリメントされる。

【0071】図5は、本発明によるさらに好ましい実施例であり、生成コンポーネントのカウンタベースでの実現を示す模式図である。この実施例では、確率的アプローチではなく決定性アプローチを使用する。このアプローチは、非常にコストのかかる乱数発生器の使用を回避する。図5において、 $i-j$ コネクションに対して充填コンポーネントにより決定されたレートを r_{ij} とする。カウンタ c_{ij} が設けられ、各時間ステップごとに r_{ij} （0と1の間の小数をとる。）だけインクリメントされる。カウンタの値は、負の値もとる浮動小数点数である。前と同様に、整数 q_{ij} は実パケットの数を保持する。スケジューラは、 $q_{ij} > 0$ または $c_{ij} \geq 1$ の場合、2部グラフ内に $i-j$ 辺があると判断する。スケジューラが $i-j$ コネクションを許可する場合、 c_{ij} は1だけデクリメントされ、実パケットがあれば（すなわち $q_{ij} > 0$ ）実パケットがルーティングされる。

【0072】[IVC. 1 シューティングブランクス・スイッチ] 図6は、本発明によるシューティングブランクス・スイッチの好ましい実施例を示す概略的構成図である。シューティングブランクス・スイッチは、 M 個の入力ポート6.11~6.14及び N 個の出力ポート6.21~6.24を有するものとする。即ち、図示し

たスイッチでは $M=4$ 及び $N=4$ である。セルは、入力ポートに到着した後、 M 個の入力ポートと N 個の出力ポートの間でルーティングされなければならない。これらのセルは、相異なるレートで到着する。各セルは、いくつかのキュー（6. 711~6. 714は一部のキューである。）のうちの1つにキューイングされる。1つのキューは、特定の入力ポートから特定の出力ポートへルーティングされるセルに対応する。例えば、キュー6. 714は、入力ポート6. 11から出力ポート6. 24へルーティングされるセルをキューイングする。同様に、キュー6. 712は、入力ポート6. 11から出力ポート6. 22へルーティングされるセルをキューイングする。従って、各入力ポートは、その入力ポートから N 個の出力ポートのそれぞれへルーティングされるセルに対応する N 個のキューを有する。こうして、このスイッチには $M \times N$ 個のキューがある。この図の場合では $4 \times 4 = 16$ 個のキューがある。

【0073】レート識別器6. 3は、最大到着レートのポートを識別する。入力ポートに対する到着レートは、その入力ポートの N 個のすべてのキューに対する到着レートの和である。同様に、出力ポートに対する到着レートは、 M 個の入力ポートのそれぞれからその出力ポートへの到着レートの和である。充填決定器6. 4は、すべてのポートに対する到着レートが等しくなるように適当なキューに対するダミーフローを決定する。ダミーセル生成器6. 5は、充填決定器によって決定された到着レートでダミーセルを生成する。スケジューラ6. 6は、キュー長を考慮しないヒュリスティックMBM法に基づいてルーティングを行う。

【0074】【IVC. 2 シューティングブランクス・スイッチを用いたネットワークシステム】図7は、シューティングブランクス・スイッチを用いたネットワークシステムの好ましい例を示す。ネットワークシステム7. 1は、いくつかのスイッチを有する。これらのスイッチのサブセットが、シューティングブランクス・スイッチ7. 21~7. 23からなる。シューティングブランクス・スイッチの好ましい実施例の構造は前のセクションIVC. 1（図6）に記載されている。

【0075】【IVD. 実験結果】注意すべき点であるが、ここに記載する実験は単なる例示であり、本発明の技術的範囲を制限するものと解釈してはならない。これらのシミュレーション実験を用いて、シューティングブランクス法が、シューティングブランクス法を使用しない従来のスケジューリングアプローチに比べて優れた結果を与えることを示す。強調されるべき点であるが、シューティングブランクス法は、任意のスケジューリングヒュリスティックと直交して組み合わせることが可

能であるが、平衡入力に最大サイズマッチングに近似する良好な方法と組み合わせるほうが良好に動作する。ここに記載する実験は、以下のようにして生成されるトラフィックパターン例を使用する。任意の時間ステップにおいて、各入力ポート i は確率 $\sum_{j=1}^N \alpha_{ij}$ で1個のセルを生成する。入力ポート i で生成されたセルの宛先が出力ポート j である確率は $\alpha_{ij} / \sum_{j=1}^N \alpha_{ij}$ である。

【0076】入力ポートの次数とは、それがマッチングされる出力ポートの数のことである。同様に、出力ポートの次数とは、それがマッチングされる入力ポートの数のことである。前述の米国特許出願及び特開平11-244648号に記載されているような低次数で不平衡のトラフィックパターンは、例とするトラフィックタイプとみなされる。このトラフィックタイプは、McKeown et al. に記載されたトラフィックパターンの一般化である。McKeownに記載されたトラフィックパターンは許容されるが、スケジューリング方法がMBMである場合にはキューが無制限に成長する（Nick McKeown, Venkat Anantharam, and Jean Walrand, "Achieving 100% throughput in an input-queued switch", In Proceedings IEEE INFOCOM '96, p.296-302, San Francisco, CA, March 1996, を参照）。今のシミュレーションの場合、半数の入力ポート及び半数の出力ポートがちょうど1個のフローをサポートし、他の入力及び出力ポートはちょうど2個のフローをサポートする。各フローは同じレートでトラフィックを生成するため、半数のポートはレート λ で負荷を受け、他の半数は 2λ で負荷を受ける。シミュレーションの開始時のフローを生成するため、2個のランダム置換を使用して、 $N!$ 個の可能性のそれぞれを等確率で選択する。重い負荷の入力ポートに対しては、置換によって定義される両方のフローを使用する。軽い負荷の入力ポートに対しては、フローを定義するのに第1の置換のみを用いる。

【0077】サイズ $N=4$ 及び $N=8$ のスイッチについて考える。完全を期するため、使用する実際のレート行列を示す。 (i, j) 成分は、入力ポート i から出力ポート j へのレートを表す。実トラフィックに対するレート行列を α_{ij} で表す。ブランク（ダミー）に対するレート行列を β_{ij} で表す。全レート行列を $\gamma_{ij} = \alpha_{ij} + \beta_{ij}$ で表す。なお、 α_{ij} は、前段落に記載したように生成され、 β_{ij} は、前述のセクションIVB. 1に記載したように α_{ij} から計算される。

【0078】サイズ $N=4$ の場合、 $\gamma_i = \alpha_i + \beta_i$ 、すなわち、

【0079】

【数1】

$$\begin{bmatrix} 0 & \lambda & \lambda & 0 \\ 0 & \lambda & 0 & \lambda \\ \lambda & 0 & \lambda & 0 \\ \lambda & 0 & 0 & \lambda \end{bmatrix} = \begin{bmatrix} 0 & 0 & \lambda & 0 \\ 0 & \lambda & 0 & 0 \\ \lambda & 0 & \lambda & 0 \\ \lambda & 0 & 0 & \lambda \end{bmatrix} + \begin{bmatrix} 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & \lambda \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

である。

【0080】サイズN=8の場合、 $\tau_8 = \alpha_8 + \beta_8$ 、すなわち、

【0081】

【数2】

$$\begin{bmatrix} 0 & \lambda & 0 & 0 & 0 & 0 & \lambda & 0 \\ \lambda & 0 & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda & \lambda & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda & \lambda & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda & 0 & \lambda & 0 \\ 0 & \lambda & 0 & 0 & 0 & 0 & 0 & \lambda \\ \lambda & 0 & 0 & 0 & 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda & 0 & \lambda \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \lambda & 0 \\ \lambda & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda & 0 & \lambda & 0 \\ 0 & \lambda & 0 & 0 & 0 & 0 & 0 & \lambda \\ \lambda & 0 & 0 & 0 & 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda & 0 & \lambda \end{bmatrix} + \begin{bmatrix} 0 & \lambda & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

である。

【0082】 $0 \leq \lambda \leq 1/2$ の範囲を考える。任意のポートに対する最大負荷は 2λ であるため、 2λ をシミュレーションのレートと呼ぶ。

【0083】実験は、特定のコンフィグレーションに対する安定性の最大レートを経験的に評価しようとするものである。ただし、コンフィグレーションは、スイッチサイズ、スケジューリング方法、及びトラフィックパターンからなる。正式な用語ではないが、入力キューが無制限に成長しない場合に、コンフィグレーションは安定であるという。安定性の経験的定義は以下の通りである：入力キューサイズの総和が、ある一定の時間ステップ数Tのうちに、ある一定の上限Qより大きく成長しない場合、コンフィグレーションは安定であると宣言される。なお、Qの増大あるいはTの減少は、一部のコンフィグレーションの実験的安定性の範囲を増大させる効果を有することがある。また、QまたはTのいずれかを増大させることは、シミュレーション時間を増大させることがある。我々の実験では、 $Q=16N^2$ 個のセルと、 $T=256N^2$ を使用する。安定性実験は、我々の安定

性基準が破られるまでにコンフィグレーションが動作可能な最大レートを決定する。このような安定性実験は非常に長くなることがあるため、離脱条件を追加する。

1. 000ステップごとにキューサイズの総和をチェックする。この総和が前回のチェック以来成長していない場合、システムは安定であると仮定される。

【0084】表1に実験結果を示す。9個のスケジューリング方法に対して、安定性のレートが、ブランクなしの4個のポート、ブランクありの4個のポート、ブランクなしの8個のポート、ブランクありの8個のポートに対してテストされている。スケジューリング方法は、3つのクラス、すなわち、MBM、高品質MBMヒューリスティックス、及び低品質MBMヒューリスティックスにグループ分けされている。ヒューリスティックスの品質とは、ブランクを追加する前のもとのトラフィックに対するMBMの挙動をどのくらいよく近似するかをいう。安定性レートが λ より低いヒューリスティックスは、ここでの考察の目的では低品質として特徴づけられる。反復法はすべて4回（4イテレーション）実行した。

【0085】主要な結果は表の第1行に示されている。N=4の場合、MBMは、ブランクなしでΛまで安定であり、ブランクありで100%まで安定である。N=8の場合、ブランクの使用は安定性レートを93%から100%に上昇させる。

【0086】次のスケジューリング方法のクラス（これは、PIM-SHAKE-4、RND-SHAKE-4、RRRND-SHAKE-4、及びSLIP-SHAKE-4を含む。）は、高品質ヒューリスティックスとみなされる。このクラスのヒューリスティックスの場合、ブランクの使用により全般的に改善が見られる。代表的な改善は、4ポートでは安定性93%から安定性9

8%への改善であり、8ポートでは安定性90%から安定性92%への改善である。なお、これらの方法が、N=4のブランクなしの場合にMBMよりも良好に動作することは、この例については説明することができるが、一般には成り立たない。これらのヒューリスティックスは、2部グラフにおいて接続している辺が多いノードのほうに偏る。これは、これらの例にとって有利である。というのは、2本の辺を有するノードは2倍の帯域をサポートする必要があるからである。

【0087】

【表1】

スケジューリング アルゴリズム	4ポート		8ポート	
	安定レート	安定レート ブランク有	安定レート	安定レート ブランク有
MBM	90%	100%	93%	10%
PIM-SHAKE-4	93%	96%	90%	92%
RND-SHAKE-4	92%	98%	90%	92%
RR-RND-SHAKE-4	93%	98%	89%	92%
SLIP-SHAKE-4	93%	100%	93%	100%
PIM-4	88%	86%	87%	86%
RND-4	88%	86%	87%	86%
RR-RND	88%	86%	88%	85%
SLIP-4	100%	80%	100%	86%

【0088】最後のスケジューリング方法のクラスは、比較的低品質のMBMヒューリスティックスである。このクラスは、PIM-4、RND-4、RR-RND、及びSLIP-4を含む。一般に、これらの方法は、ブランクを使用するとわずかにパフォーマンスが悪くなり、いずれのスイッチサイズの場合でも安定性レートが88%から86%に低下する。SLIP-4は例外である。SLIP-4の場合の100%のスループットは、我々のトラフィックフロー例が与えられれば、予測することができる。優先度ホイールが完全に同期することにより、100%のスループットが可能となる。しかし、SLIP-4は、この例における平衡トラフィックでは問題がある。一般に、SLIPは、多くの低次数トラフィックパターンに対して高品質の解を達成するのは困難であることを示すことができる（前述の米国特許出願及び特願平11-244648号を参照）。この実験においてブランクなしのSLIPの良好なパフォーマンスは、単純なトラフィックパターンによる見かけ上の結果である。

【0089】本発明の他の変形例は、当業者には以上の説明から明らかである。本発明のいくつかの実施例についてのみ具体的に説明したが、本発明の技術思想及び技術的範囲から離れることなく、さまざまな変形が可能であることは明らかである。

【0090】

【発明の効果】以上詳細に説明したように、本発明によるスケジューリング方法によれば、まず、1つのポート

を最大のセル到着レートを有するとして識別し、次に全てのポートに対する到着レートがほぼ等しくなるように、最大のセル到着レートを有するとして識別されたポート以外のすべてのポートにダミーセルを追加する。こうすることで、MWBM（最大重み2部マッチング）スケジューラの優れた安定性を達成することができるとともに、より単純なMBM（最大サイズ2部マッチング）またはMBMベースのヒューリスティックスを実行することができる。

【図面の簡単な説明】

【図1】スケジューリング問題の2部グラフを説明するための模式図である。

【図2】仮想出力キューイング（VOQ）をサポートするN=4のクロスバスイッチの模式的構成図である。

【図3】不平衡フローのスイッチの一例を示す模式図である。

【図4】本発明の一実施形態における充填コンポーネントによるダミーフローの生成後の、図3のトラフィックの新しいレートを示す模式図である。

【図5】本発明によるさらに好ましい実施形態であり、生成コンポーネントのカウンタベースでの実現を示す模式図である。

【図6】本発明によるシューティングブランク・スイッチの好ましい実施例を示す概略的構成図である。

【図7】本発明によるシューティングブランク・スイッチを用いたネットワークシステムの好ましい実施例を示す模式図である。

【符号の説明】

6. 11~6. 14 入力ポート
6. 21~6. 24 出力ポート
6. 3 レート識別器
6. 4 充填決定器
6. 5 デミセル生成器

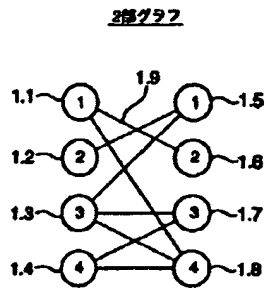
6. 6 スケジューラ

6. 711~6. 714 キュー

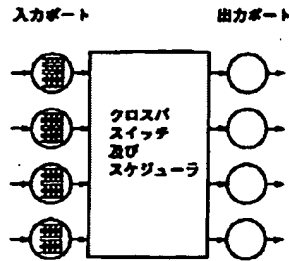
7. 1 ネットワークシステム

7. 21~7. 23 シューティングブランクスイッチ
チ

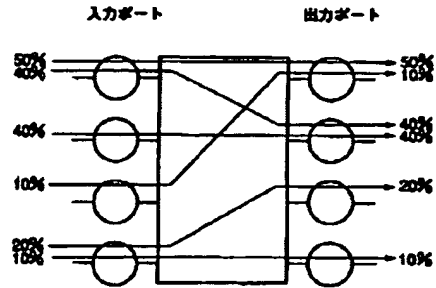
【図1】



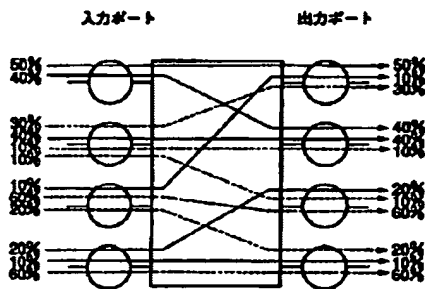
【図2】



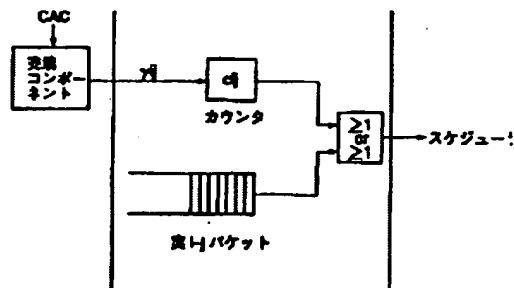
【図3】



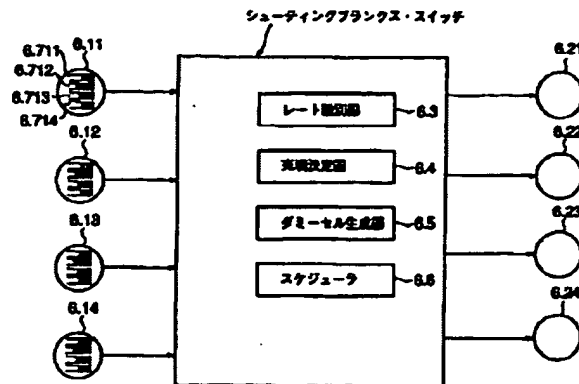
【図4】



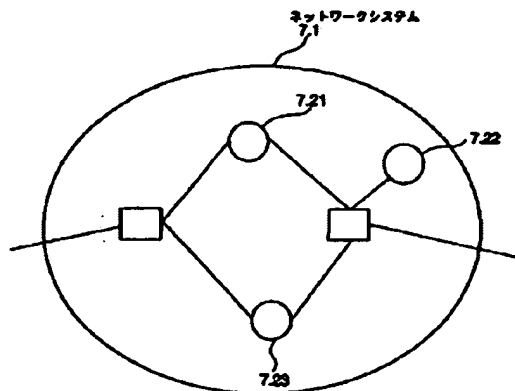
【図5】



【図6】



【図7】



【手続補正書】

【提出日】平成11年9月1日(1999. 9. 1)

【手続補正1】

【補正対象書類名】明細書

【補正対象項目名】発明の名称

【補正方法】変更

【補正内容】

【発明の名称】 キューイングシステムにおけるスケジューリング方法及びシューティングブランクス・スイッチ

フロントページの続き

(72)発明者 スタッフロス・コロオパロス
アメリカ合衆国、ニュージャージー
08540 プリンストン、4 インディペン
デンス ウエイ、エヌ・イー・シー・リサ
ーチ・インスティテュートウ・インク内

(72)発明者 サティシ・ラオ
アメリカ合衆国、ニュージャージー
08540 プリンストン、4 インディペン
デンス ウエイ、エヌ・イー・シー・リサ
ーチ・インスティテュートウ・インク内

(註6) 100-295247 (P2000-29818

Fターム(参考) SK030 GA03 HA10 HB09 HB17 KA03
KA13 KX12 KX18 KX29 LA03
LB07 LE05 MA13
9A001 CC03 CC07 FZ05 HH34 KK56
LL09